An Efficient Data Management Framework for Puerto Rico Testsite for Exploring Contamination Threats (PROTECT)

Shi Dong[1], Zlatan Feric[1], Leiming Yu[1], David Kaeli[1], John Meeker[3], Ingrid Y. Padilla[4],
Jose Cordero[5], Carmen Velez Vega[6], Zaira Rosario[6], Akram Alshawabkeh[2]
[1]Dept .of Electrical and Computer Engineering, Northeastern University
[2]Dept. of Civil and Environmental Engineering, Northeastern University
[3]University of Michigan [4]University of Puerto Rico at Mayaguez [5]University of Georgia
[6]Graduate School of Public Health, University of Puerto Rico Medical Campus

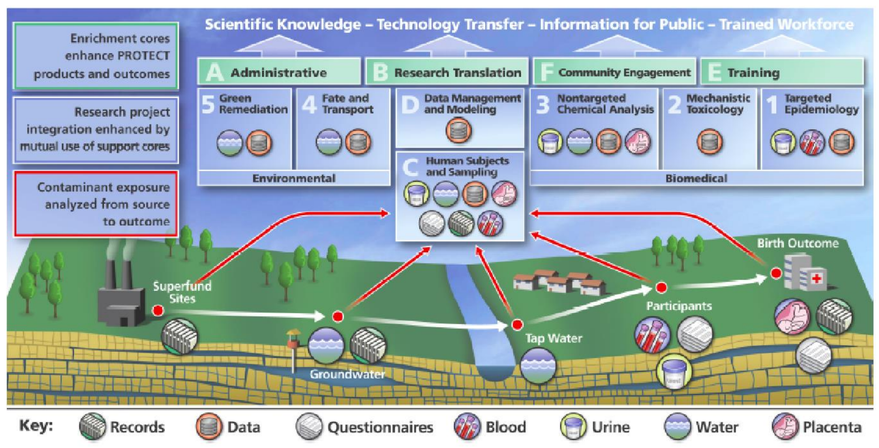
Abstract

In the era of Big Data, both the volume and complexity of data grows rapidly. However, we face many challenges when trying to manage this scale of data, given that different project domains require tailored methods of management. In this poster, we present:

• An efficient data management framework for the NIEHS-supported Puerto Rico Testsite for Exploring Contamination Threats (PROTECT) Center.

• A series of associated workflows, supporting the tasks of data import, data cleaning, and secure transmission of privacy-sensitive PROTECT data, while enabling online data inquiry, visualization, and data processing to support data analytics.

PROTECT



Human Subject Data

- 14 questionnaires
- 1,800 participants
- 3,552 total fields per participant
- Anticipated total fields >6M

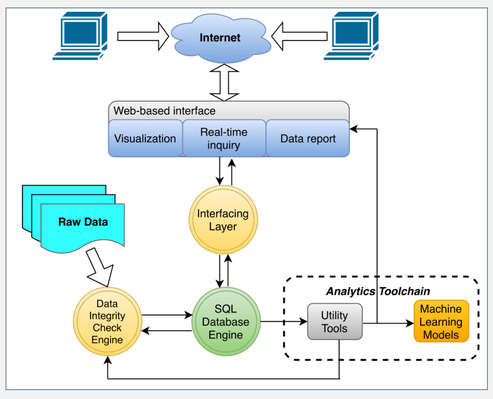
Environmental Data

- Archival/historical data from government resources
- Field samples from wells, rivers and tap water
- Measuring Phthalate, CVOC, Water level, Discharge and Meteorological Data

Biological Data

- Testing participant urine
- 14 analytes for pesticides/participant
- 18 analytes for trace metals/participants
- 19 analytes for biological chemicals

Data Management Framework

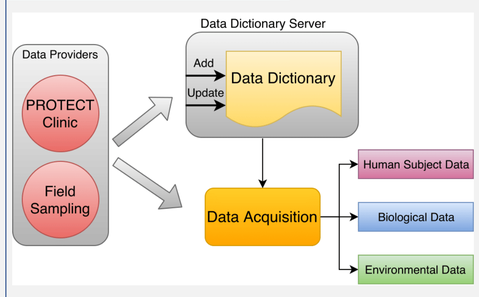


- A web-based interface for data visualization, data query handling, and data report generating
- A data cleaning engine to ensure the data integrity
- A database engine to maintain and index all of the data
- A set of utility tools for multiple purposes, including workflow facilitation, data statistics and visualization, and data processing for machine learning studies

Workflows

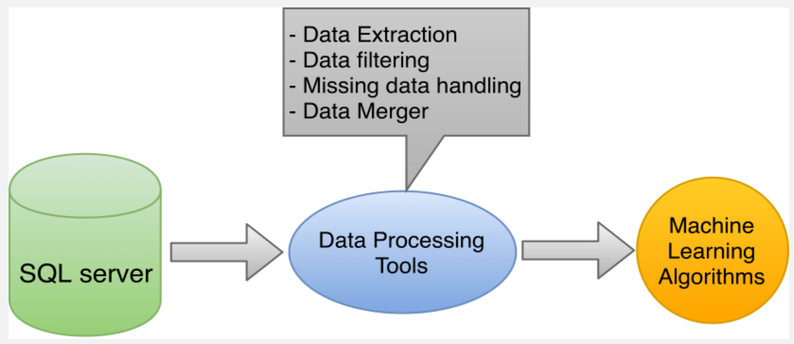
Workflows are defined to serve a range of needs in the PROTECT Center.

Data Acquisition



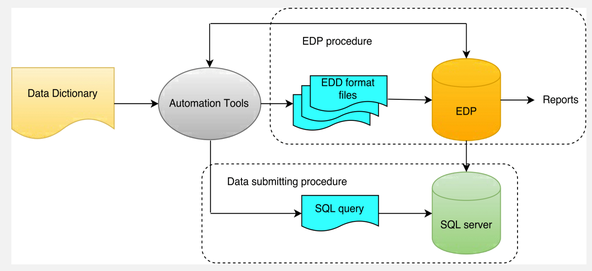
- Carefully design data dictionaries for each data entries
- Maintain and keep the data dictionaries up-to-date
- Collect data from different sources according to the data dictionaries

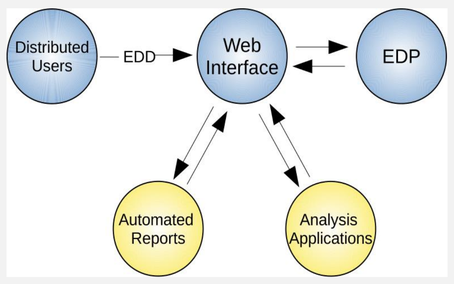
Data Processing

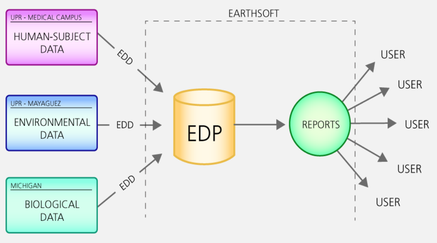


- An analytics tool chain is developed to process data directly from the database
- The data can be fed to Machine Learning models for data mining/data analytics

Data Cleaning , Incorporation, and Sharing



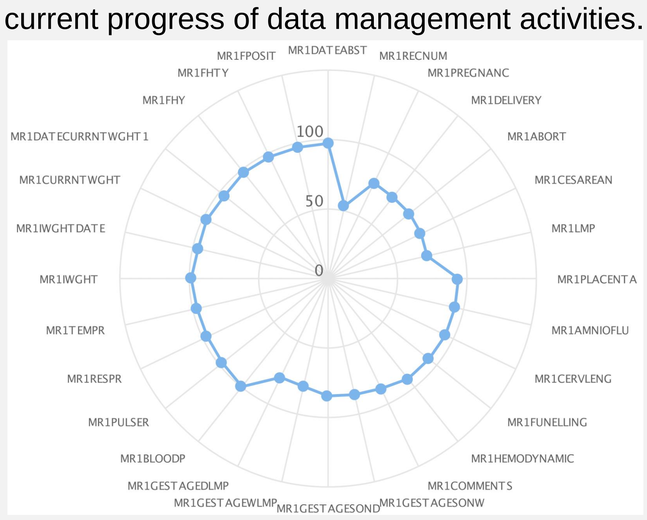


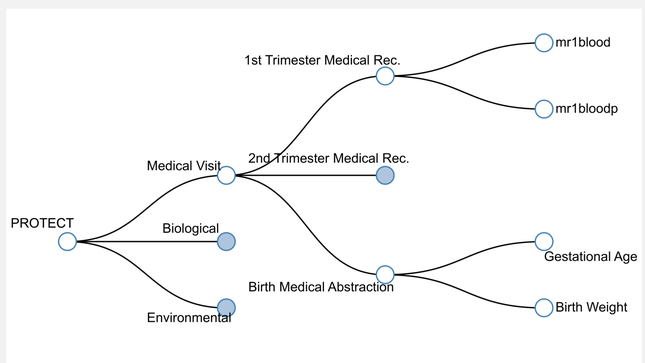


- Data cleaning is conducted to guarantee data integrity through the EDP toolset
- Web-based interface services users from distributed locations
- In-house tools are developed to facilitate the workflow

Data Visualization

Visually inspecting data can significantly improve a researcher's ability to quickly understand their data. We also use visualization to track the status of the data management system, delivering clear messages of current progress of data management activities.





Average Phenol Concentration in Urine

Name	24-DCP	25-DCP	B-PB	BP-3	BPA	BPF	BPS	E-PB
Camuy Health Services	12.24	460.86	0.39	164.47	3.51	0.31	0.22	83.18
Centros Integrados de servicios de Salud Lares	1.30	27.58	11.20	748.62	3.93			
Manati Medical Center	3.21	94.50	4.49	416.76	4.20	0.28	0.66	37.93
Morovis Community Health Center	4.44	144.00	6.64	196.55	3.66	0.26	0.20	12.40
Prymed Ciales	1.94	56.16	3.14	192.75	5.09	0.30	1.70	13.00
Average	3.50	108.12	4.54	364.28	4.21	0.28	0.56	40.01

Conclusion

- We present an efficient data management framework that leverages both existing tools and our own customized tools to effectively manage the data of the NIEHS PROTECT Center.
- We present our general workflows, architecture, and toolsets, which streamline workflow and provide solutions for data processing for machine learning algorithms

Reference

[1] IBM, P. Zikopoulos, and C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 1st ed. McGraw-Hill Osborne Media, 2011.
[2] EarthSoft, "EQuIS Professional," http://www.earthsoft.com, 2017 (accessed October 1, 2017).
[3] K. Delaney, Inside Microsoft SQL Server 2000. Redmond, WA, USA: Microsoft Press, 2000.
[4] Wen Jiang, "PyPyODBC 1.3.5," https://pypi.python.org/pypi/pypyodbc, 2017 (accessed October 1, 2017).
[5] Mike Bostock, "D3.js," https://d3js.org/, 2016.
[6] X. Li, L. Yu, D. Kaeli, Y. Yao, P. Wang, R. Giese, and A. Alshawabkeh, "Big Data Analysis on Puerto Rico Testsite for Exploring Contamination Threats," ALLDATA, 2015.

Acknowledgment

This work is supported in part by NIEHS Superfund Research Program award P42P017198.